Paper code: CAV05081

AN ILLUSTRATIVE EXAMPLE OF THE BENEFITS OF USING A RASCH ANALYSIS IN AN EXPERIMENTAL DESIGN INVESTIGATION

Robert F. Cavanagh
Curtin University of Technology

David B. Kent
Curtin University of Technology

Joseph T. Romanoski
Curtin University of Technology

Paper presented to the Assessment and Measurement Special Interest Group at the 2005 Annual Conference of the Australian Association for Research in Education: Sydney

Address correspondence to Dr Rob Cavanagh

Curtin University of Technology

Department of Education

GPO Box U1987

Western Australia 6845

Email: R.Cavanagh@curtin.edu.au

Abstract

The study compared the consequences of choosing to use either deterministic or probabilistic data analyses in an experimental investigation. The empirical research from which the data for these analyses was drawn applied computer assisted language learning (CALL) as the treatment in a one-group pretest-posttest design. The empirical investigation concerned the effect of CALL on Korean university students' ability to correctly identify the meaning of loanwords - native vernacular that originated in non-native languages and is now part of the native vernacular. The empirical investigation is explained and the design of experimental research and analysis of experimental research data are discussed. Stochastic and deterministic measurement models are then examined followed by the application of these models to analyse data from the empirical investigation. Data analyses included a paired sample *t*-test, one-way analysis of variance (ANOVA) and a Rasch Unidimensional Measurement Model (RUMM) analysis of differential item functioning (DIF). The analyses and their respective results are assayed in terms of capacity to inform hypothesis confirmation. Parametric tests using non-interval data (raw test scores) were shown to be less sensitive than the RUMM DIF analysis of Rasch Model transformed scores when estimating the differences between the pretest and posttest data.

AN ILLUSTRATIVE EXAMPLE OF THE BENEFITS OF USING A RASCH ANALYSIS IN AN EXPERIMENTAL DESIGN INVESTIGATION

Overview

This report used data from an experimental design investigation to examine the appropriateness of deterministic and probabilistic data analysis techniques. The report commences with a brief discussion of the empirical investigation leading on to a general examination of experimental research and the analysis of data from experimental research. Then, data from the empirical investigation is analysed from the differing perspectives of deterministic and probabilistic approaches.

The empirical research

The aim of the empirical research was to examine the educational effectiveness of using the English inherent within the native vernacular of Korean EFL students for the development of English language linguistic competence. Specifically, to investigate the utility of using computer-assisted language learning (CALL) homework within a mandatory university English program.

It was envisioned that linguistic competence of the Korean EFL student could be effectively stimulated through the use of loan terminology, or more precisely the English that now forms a part of the learners' native vernacular (Kent, 2000). Such terminology is intrinsically tied to the original culture (Taylor & Taylor, 1995), and it is believed that such nativised lexical elements can function as cross-linguistic mnemonic keys for phrases and vocabulary learnt in the target language (Daulton, 1999a). Korean learners, even at the advanced level, inappropriately and continually misuse these terms in English conversation (Kent, 2000). The use and misuse of this terminology leads to misunderstanding and confusion in the language learning process (Sheperd, 1996). Students are also unaware of the contextual use of such vocabulary and also of the differences in meaning in Korean from the English language source (Shaffer, 1999).

Although usefulness of the English inherent within the native vernacular is recognized as a learning resource for Japanese EFL students (Daulton, 2003; Shepherd, 1996; Simon-Maeda, 1997), and their Korean counterparts (Kent, 1996; Shaffer, 1999), examination of the effectiveness of the practical application of such terminology for second language acquisition has not occurred in Korea. In Japan, Daulton (1998) highlights the limited research available that illustrates the positive effect of loanwords on English vocabulary acquisition for Japanese EFL students at all levels. Unfortunately, no such experimental studies are available involving Korean learners. Daulton (1998) further indicates that 20,000 English word forms are used in Japanese, and comprise 10% of the language. In more recent research Daulton (2003), examining vocabulary lists, also indicates that of the 3,000 most-frequent word families in English, as found in the British National Corpus (BNC 3000), 45.5% correspond to common Japanese loanwords, and in using the West General Service List (GSL) uncovered a 38% correspondence. The correspondence of Korean loan terms on such vocabulary lists is yet to be determined. In addition, there is difficulty in uncovering data granting an exact indication as to the proportion of the use of English in the native Korean vernacular. However, Kent (1996) does provide a dictionary consisting of 1,167 everyday English and European loan expressions, and as Taylor and Taylor (1995, p. 197) note: "What proportion of the present-day Korean vocabulary is European in origin? I have not found any data on this question, but in the Dictionary of Terms for Current Affairs [Dong-Ah, 1987] 28% are European, mostly English, loanwords".

In the past, extensive use of the native language in EFL has occurred through such largely criticized methodologies as Grammar-Translation, this has since given way to approaches like the English-only policy of the Communicative Language Teaching Approach. The theoretical basis underlying the communicative approach stems from the view that second languages are learnt in a similar fashion to first languages, and that language is a system for expression of meaning; the

primary function of language is for interaction and communication; and, the structure of language reflects its functional and appropriate communicative use. In this regard, the ability to be understood is more important than the grammatical correctness of the linguistic message.

In Korea, government reforms for the development of information communication technology within higher education focus firmly on administrative use, infrastructure development, and the promotion of research (K.E.R.I.S., 2001), in contrast to the grade school setting, where computer-based initiatives are used to assist learning. While in other nations, "… we have been witnessing a steady increase in online course delivery by tertiary education providers" (Elgort, Marshall & Mitchell, 2003, p. 1), there is lack of focus on the development of higher education e-learning initiatives in Korea. Further, "textbooks are preferred in university-level curriculum" so, "in spite of the availability and accessibility of computers and the internet today, the integration of web-technology into the curricula of Korean universities has not found widespread acceptance" (Min, Kim & Jung, 2000, p.120). As a result, "… in many respects English-teaching and language learning methods in Korea have not yet caught up with the times. Centuries-old methods of dealing with both teaching and learning languages are still closely adhered to" (Shaffer, 2001, p. 1). Although there are "very few cases of using CALL in regular [university] English programs" (Lee & Yang, 2002), numerous Korean educators employ CALL at the tertiary level within individual EFL classes. These educators use chat programs or e-mail for computer mediated communication (S. Y. Kim, 1999 & 2002), the internet for reading activities and presentations (Kang, 2000), or multimedia material like language learning software for improving linguistic skills (Keem, 2000).

Implicit in the increased use of CALL for EFL is the assumption that gains in EFL competence will result from application of CALL strategies. This assumption was tested by administering a CALL instructional program to a sample of Korean university students and then testing the effect on their ability to correctly identify the English meaning of Korean loanwords.

The core linguistic content of the CALL treatment was selected from a vocabulary of direct loanwords, false-cognates, hybrid terms and substitution terms. Direct loanwords are those that contain the same meaning in both the original and borrowing language – e.g. *juice* and *stress*. False cognates are words that maintain the original pronunciation but possess different meanings in the borrowing language – e.g. *cunning* meaning *deceptive* or *exhibiting ingenuity* in English but meaning *cheating* in Korean. Hybrid terms are words formed by a mixture of the original and borrowing languages – e.g. *com-maeng* meaning *computer illiterate*. Substitution terms are words that are now commonly used in place of the native term in the borrowing language – e.g. *a-reu bai-teu* (from German) meaning *part-time job*. The CALL treatment presented the vocabulary to be learnt in three different activities. The first activity provided students with classification or sorting-based tasks. The second activity of the treatment required students to attempt multiple-choice selection-based tasks. The third activity involved students in clue-type matching-based exercises. In each language learning activity, context sensitive feedback was provided to students so they repetitively reviewed the material and also maintained control over the pace and sequence of activities.

The effect of the treatment on student ability to correctly identify the English meaning of Korean loanwords utilised administration of a pre-treatment and post-treatment test - a one-group pretest - posttest experimental design. The test was multiple-choice with students asked to identify the correct English meaning of a particular Korean loanword from four alternative statements of English meaning. The test was developed by an iterative process of item writing, trialling with 50 students, and examination of descriptive statistics. Student raw scores (N=50) were used as an estimate of student ability and the frequency of incorrect answers for each item as an estimate of item difficulty. 40 items were selected that produced a range of student scores that 'matched' the range of item difficulties. Given the limitations of this process, it was anticipated the test would require refinement when the research sample was tested.

Experimental research

Research designs

Experimental research involves manipulation of a variable(s) and tests for the effect of this manipulation or treatment on a dependent variable(s). It is a powerful method that can provide strong evidence for confirmation of hypothesised cause and effect relationships. There are multiple experimental research designs varying in how threats to internal validity are controlled and correspondingly, in the complexity of the data analysis procedures.

Fraenkel and Wallen (2004) classify experimental research designs as 'weak', 'true' or 'quasi-experimental' depending on the extent to which a design controls for the effect of threats to internal validity. Weak designs include: the one-shot case study (treatment followed by observation); the one-group pretest-posttest design (pretest, treatment then posttest); the static group comparison design (observations of two existing groups are compared); and the static-group pretest-posttest design (two existing groups are administered a pretest, treatment then posttest). The key difference between these designs and the true experimental designs concerns the assignment of subjects to treatment groups by random sampling or random sampling with matching of subjects according to certain variables(s). Alternatively, quasi-experimental designs apply non-random assignment but reduce threats to internal validity by techniques such as: matching subjects; a common treatment is administered at different times to different groups (counter-balanced designs); repeated measurements over time (time-series designs); and inclusion of additional or moderator variables (factorial designs).

These designs have differing effectiveness in controlling for internal validly threats such as subject characteristics, mortality, location, instrument decay, data collector characteristics, data collector bias, testing, history, maturation, attitudinal, regression and implementation (Fraenkel & Wallen, 2004). Notwithstanding these threats, experimental designs enable a high degree of control to be exercised over extraneous variables by techniques including random assignment of subjects to experiment and controls groups, holding certain variables constant, building variables into the design, matching to produce equivalent comparison groups, and analysis of covariance to statistically equate groups and adjust posttest scores according to data from a pretest or other variable(s).

Of particular interest in this paper is the matter of statistical analyses and how covariance in data from experimental and extraneous variables is dealt with.

Data analyses

Application of statistical analyses enables answering of questions concerning inferences about the data. One means for answering such questions is using the *t*-test to determine whether the difference between the mean scores from two groups is significant. This is a parametric technique since it is based on the assumptions that the characteristic of interest is normally distributed within a population and also that the distribution of the differences between the group means is equal to the difference between the means of the two 'populations' from which the groups were drawn . Similarly, estimating the variance both within and between groups using analysis of variance (ANOVA) is also a parametric technique. In comparison to the *t*-test, ANOVA can test for differences in the means for more than two groups and for more than one independent variable. Also, the strength of the effect of the treatment can be estimated by calculating the eta^2 statistic. Additionally, analysis of covariance (ANCOVA), another parametric analysis, can be applied to adjust scores to compensate for differences between groups in data from a related variable. When applying the one-group pretest-posttest experimental design, ANCOVA can adjust the posttest scores to compensate for differences between groups revealed by the pretest scores.

At the research design stage, controlling for internal validity threats often does not include consideration of instrumentation and whether the instrument is a measure of the characteristic of interest (see preceding list of threats to internal validity) - the test scores are often assumed to be a

measure of the subject's ability. A concise presentation of the postulates or requirements for measurement is found in the book *Rating Scale Analysis* (Wright and Masters, 1982). These requirements can be rephrased to describe those features which a number must manifest in order to be considered a measurement:

- Unidimensionality - the reduction of experience to a one dimensional abstraction (height, weight, intelligence);
- Qualification - more or less comparisons among persons, items, etc. (taller or smaller, heavier or lighter, brighter or duller);
- Quantification - a unit determined by a process which can be repeated without modification over the range of the variable (feet, inches, pounds, logits); and
- Linearity - the idea of linear magnitude inherent in positioning objects along a line by some device or instrument (tape measure, scale).

With specific regard to parametric tests and their suitability for different types of data, Fraenkel and Wallen (2004, p. 241) noted:

> "It turns out that in most cases parametric techniques are most appropriate for interval data, while nonparametric techniques are most appropriate for ordinal and nominal data. Researchers rarely know for certain whether their data justify the assumption that interval scales have actually been used."

Likewise, with regard to the use of raw scores with ANOVA, Romanoski and Douglas (2002, pp. 233-234) asserted that:

> "For the most part, test scores are considered to be 'measures' of student abilities. The main issue at stake is that, for the best part of the twentieth century, the validity of using certain statistical techniques in the analysis of test scores, particularly the use of the statistical technique of analysis of variance (ANOVA), has rarely been questioned. The possibility that the number of correct answers on a test might not constitute a 'measure' for the purposes of statistical analysis has been largely overlooked." (p. 233)

> "… ANOVA is extensively utilized in studies in which an independent variable is routinely examined for interactions with another independent variable. The discovery of spurious interaction effects produced under ANOVA of raw scores (Embretson, 1993, 1996), when Rasch transformations of the raw scores could easily have been employed, calls this use into question. Embretson found that when untransformed raw scores are subjected to multi-factor ANOVA spurious interaction effects between the independent variables regularly occur. Since the interaction effect, in many cases, reflects the major research hypothesis, this finding should be of concern to researchers and statisticians alike." (p. 234)

The point here is that the application of parametric tests to data which are non-interval may well render the test results invalid and it is important for the researcher to confirm that the data can be plotted on an interval-level scale. This is a measurement issue and it can be understood by examining 'stochastic' and 'deterministic' measurement models.

In stochastic measurement models, probability refers to the total score for a subject predicting with varying degrees of certainty the correct answering of test items - the relationship between total score and a person's ability to correctly answer questions is probabilistic (Bond & Fox, 2001, pp. 233-234). In deterministic measurement models, the relation between the observed responses and person ability is explicated as a causal pattern Bond & Fox, 2001 p. 229) - raw scores are often taken as a measure of person ability (Bond & Fox, 2001, p. 2). It follows that parametric tests typically used in experimental research are deterministic since the subject's score is assumed to be a measure of the subject's ability. It should be noted that the term 'probability' is applied in parametric analyses and this use should not be confused with its use in probabilistic analyses. In parametric analyses, probability basically refers to the level of certainty when comparing sample

means or comparing a sample mean with a population mean. That is, the estimation of probability assumes normal distributions of group (sample) scores, population scores, differences between group scores, and the standard error of the difference between group means.

The Rasch Model is a stochastic model. Rasch measurement takes into account two parameters - test item difficulty and person ability. While these parameters are assumed interdependent, separation between the parameters is also assumed - "mutual conformity and separability" (Rasch, 1960, pp. 121-125). For example, the items (questions) within a test are hierarchically ordered in terms of their difficulty and concurrently, persons are hierarchically ordered in terms of their ability. The separation is achieved by using a probabilistic approach in which a person's raw score on a test is converted into a success-to-failure ratio and then into the logarithmic odds that the person will correctly answer the items - a logit. When this is estimated for all persons, the logits can be plotted on one scale. The items within the test can be treated in a similar manner by examining the proportion of items answered incorrectly and then converting this ratio into the logarithmic odds of the item being incorrectly answered - a logit. When this is estimated for all items, the logits can be plotted on one scale. A person's logit score can then be used as an estimate of that person's ability and the item logit score can then be used as an estimate of that item's difficulty. Since person ability was estimated from the proportion of correct answers and item difficulty from the proportion of persons with incorrect answers, both these estimates are related and the relationship between them can be expressed as a mathematical equation - the Rasch Simple Logistic Model. This Rasch model is used to calculate person abilities, to calculate item difficulties, and then to plot the person abilities and item difficulties on the same scale. According to the model, the probability of a person being successful on a given item is an exponential function of the difference between that person's ability and the difficulty of the item.

With regard to the use of pre and posttests, accounting for the effects of covariates and controlling extraneous variables in experimental research designs, using the Rasch Simple Logistic Model to plot both pre and posttest scores on the same interval-level scale has advantages over deterministic analytic techniques. Data pre intervention and post intervention can be plotted on one interval-level scale. Also, by calibrating person ability against item difficulty, the influence of extraneous variables is reduced. Further, by measuring person ability pre and posttest in the same units on one scale, the pre and posttest transformed scores are accurate measures and can be statistically compared with a high level of confidence. This attribute of the Rasch approach can be compared with the deterministic approach that requires the data to be a measure, but does not test this requirement, and then proceeds with analyses that assume the data are a measure.

The following sections of this report examine data from the one-group pretest - posttest empirical investigation from the differing perspectives of deterministic and stochastic data analysis techniques.

The research problem and objectives

The research problem centred on the comparative degrees of accuracy of results obtained by applying *t*-tests and ANOVA to raw data from an experimental study in comparison to ANOVA of data that have been calibrated through a Rasch model transformation. The contingent objectives were to:

1. Test for the effect of the treatment on the dependent variable (the test) raw scores using a paired-sample *t*-test and ANOVA including estimation of eta^2;
2. Calibrate the dependent variable scores against the difficulty of the items in the test using the Rasch Unidimensional Measurement Model (RUMM) computer program (Andrich, Sheridan, Lyne & Luo, 2000); and

3. Use RUMM ANOVA to test for the effect of the treatment on the calibrated pre and post-treatment test scores.

Methods

Data were obtained from administering a 40 item multiple choice test to a convenience sample of 108 Korean university students studying five courses of study (childhood education, electronics, elementary education, pharmacy and occupational therapy) before and after completion of a one-semester CALL homework program. The differences between the pre and post-treatment raw scores were initially examined by conducting a paired-sample *t*-test and one-way ANOVA including estimation of effect size (eta^2) using SPSS 12.01 for Windows (SPSS, 2003).

The psychometric properties of combined pre and post-treatment data were also examined by RUMM analyses (Individual Item Fit Statistics) and the test was refined by removing items with poor fit of the data to the Rasch measurement model. A second RUMM analysis was conducted of data from the retained items to calibrate the scores and to use RUMM ANOVA to test for differences between pre and post-treatment transformed scores. Apart from using transformed scores, a second feature of this analysis was that the scores as expected from applying the measurement model were considered in conjunction with a measure of the student's ability (a logit - the logarithmic odds that the student will correctly answer the items). Three class intervals of student ability were estimated by RUMM and the effect of the treatment on these three 'classes' of students was calculated. In this way, the effect of the treatment was calculated separately for students of differing ability and for the treatment to be deemed as affecting the score, it needed to significantly affect the scores for all the three 'classes' of students in the same direction.

Results and discussion

Parametric analyses

Table 1 presents the results of the t-test of aggregated pre and post-treatment raw scores (40 items). The low probability value (Sig. [2-tailed]) of 0.0000 indicates that the difference between the pre and post-treatment aggregated raw scores is statistically significant and this finding would confirm the hypothesis that the treatment has positively affected student performance on the test.

Table 1

Paired sample t-test of pre and posttest scores (40 items; N = 108)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	VAR00001	23.85	108.00	4.73	0.46
	VAR00002	30.62	108.00	5.41	0.52

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	VAR00001-VAR00002	-6.77	5.03	0.48	-7.73	-5.81	-13.99	107	0.000

Table 2 presents the results of ANOVA between pre and post-treatment aggregated raw scores (40 items) including estimation of the eta^2 statistic. As was the case with the t-test results, the very low probability value provides evidence to confirm the hypothesis that the treatment has positively affected student performance on the test. Additionally, the eta^2 value of 0.31shows a 31% difference in the variance between the pre and post raw scores. This finding suggests the treatment has had a strong effect - eta^2 values >0.15 are considered high in behavioural science research (Kiess, 1996).

Table 2
ANOVA of pre and posttest total raw scores (40 items; N = 108)

test	Mean	N	Std. Deviation
Pretest	23.8	108	4.73
Posttest	30.6	108	5.41
Total	27.2	216	6.10

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.	Eta^2
score * test	Between Groups	(Combined)	2473.8	1	2473.9	95.7	0.000	0.31
	Within Groups		5531.0	214	25.8			
	Total		8004.9	215				

Table 3 presents the results of ANOVA between the pre and post-treatment raw scores for the 40 individual items. The difference between the scores for 20 of the 40 items was statistically significant (Sig. <0.05). Interestingly, two of the items elicited scores that were lower on the posttest than the pretest (items 4 and 22) although the differences were not statistically significant. This finding raises questions about what the two items were 'measuring'. For example, were they eliciting responses on the same student ability as were the other 38 items? The values of the eta^2 statistic show a strong effect of the independent variable on the dependent variable for five items (eta^2 >0.15) and varying levels of effect on the other 35 items.

Table 3
ANOVA of pre and posttest raw scores for each item (40 items; N = 108)

	Pretest correct responses/108	Posttest correct responses/108	F	Sig.	Eta^2
1	86	97	4.38	0.038	0.02
2	103	105	0.52	0.473	0.00
3	102	105	1.04	0.309	0.00
4	67	60	0.93	0.335	0.00
5	66	92	17.04	0.000 [a]	0.07
6	42	80	30.83	0.000 [a]	0.13
7	80	101	16.01	0.000 [a]	0.07
8	82	98	8.80	0.003 [a]	0.04
9	71	86	5.33	0.022	0.02
10	83	95	4.65	0.032	0.02
11	56	78	9.86	0.002 [a]	0.04
12	80	95	6.93	0.009	0.03
13	10	54	53.17	0.000 [a]	0.20 [b]
14	91	92	0.04	0.851	0.00
15	92	102	5.13	0.024	0.02
16	48	73	12.30	0.001	0.05
17	62	78	5.28	0.023	0.02
18	93	97	0.70	0.405	0.00
19	104	105	0.15	0.702	0.00

Table 3 continued

20	6	45	47.21	0.000 [a]	0.18 [b]
21	93	96	0.38	0.539	0.00
22	77	72	0.54	0.464	0.00
23	104	107	1.84	0.176	0.01
24	81	89	1.77	0.185	0.01
25	46	70	11.18	0.001 [a]	0.05
26	42	60	6.13	0.014 [a]	0.03
27	77	87	2.54	0.113	0.01
28	73	89	6.45	0.012 [a]	0.03
29	20	58	33.16	0.000 [a]	0.13
30	47	76	16.98	0.000 [a]	0.07
31	96	100	0.88	0.350	0.00
32	54	81	15.29	0.000 [a]	0.07
33	70	93	13.96	0.000	0.06
34	21	71	60.06	0.000 [a]	0.22 [b]
35	24	69	46.03	0.000 [a]	0.18 [b]
36	54	78	11.73	0.001 [a]	0.05
37	46	72	13.29	0.000 [a]	0.06
38	3	26	23.13	0.000 [a]	0.10
39	24	68	43.74	0.000 [a]	0.17 [b]
40	100	107	5.78	0.017 [a]	0.03

[a] denotes $p<0.05$ – 20 items

[b] denotes eta^2 >0.15 – strong effect of independent variable on the dependent variable – 5 items

At this juncture in presenting the data analysis results, the internal reliability of the 40-item scale for the sample investigated requires questioning as does the content validity of the scale. Post-hoc analysis of scale reliability was conducted. Cronbach's Alpha was 0.71 for the 40 items in the pre-treatment data, 0.81 for the 40 items in the post-treatment data. Also deleting items, specifically items 4 and 22, had minimal effect on scale variance. With regard to content validity, the reason for developing and applying the scale was to test for the effect of the treatment on the dependent variable. Some items appear to have this capacity as evidenced by a large difference between the pre and post-treatment scores and others do not (small difference between pre and post). The nub of this problem likely concerns the capacity of the 40-item scale to be a measure of the student ability under investigation, both before and after the treatment was administered. The pre and post-treatment data on the numbers of students who were correct in their responses to the two administrations of the test (see Table 3) signalled these problems, but more importantly, draws attention to the need to conjointly take into account both the characteristics of the students (their ability to correctly answer the questions) and also the characteristics of the test items (the level of difficulty of the items). As was noted earlier in this report, the Rasch model is a stochastic model and Rasch measurement takes into account the two parameters of person ability and item difficulty. Accordingly, the following data analyses apply Rasch model techniques.

Probabilistic analyses

A Rasch analysis of combined data from the pre and post-treatment tests calibrated item difficulty against student ability and estimated the fit of individual items to the Rasch model. Table 4 presents the results of this analysis including the item locations in logits (the logarithmic odds of the item not being correctly answered), the residuals for each item (the difference between the actual score and the score expected by application of the measurement model), and chi-square statistics (evidence of item data to measurement model fit). Since these results were derived from

conjoint analysis of student ability and item difficulty, there is evidence that some items are not measuring the student ability under investigation as well as other items. For example, items 4, 22 and 36 elicited data with high residuals showing these items were less accurate indicators of the student ability. Also, items 4, 20, 22, 34, 35, 36 and 38 produced data with poor fit to the measurement model (high chi-square value and low probability). These findings suggest that the capacity of the test to elicit student and item data that can both be accurately plotted on one interval level scale (a requirement of objective measurement) might be improved by deleting the seven items.

Table 4
Individual item fit – serial order of 40 items (N = 216)

Item	Location	SE	Residual	DegFree	DatPts	Chi Sq	Prob	degF
1	-0.81	0.19	0.07	209.63	216	0.92	0.63	2
2	-2.46	0.35	-0.49	209.63	216	0.58	0.74	2
3	-2.26	0.33	-0.18	209.63	216	0.36	0.83	2
4	0.74	0.15	4.18 [a]	209.63	216	17.40	0.00 [b]	2
5	-0.06	0.16	-0.58	209.63	216	1.60	0.45	2
6	0.81	0.15	-2.09	209.63	216	3.74	0.15	2
7	-0.83	0.19	-1.41	209.63	216	5.16	0.07	2
8	-0.68	0.19	0.12	209.63	216	0.60	0.73	2
9	-0.01	0.16	0.15	209.63	216	0.57	0.75	2
10	-0.66	0.18	-0.90	209.63	216	1.41	0.49	2
11	0.56	0.15	0.91	209.63	216	2.80	0.24	2
12	-0.54	0.18	-0.09	209.63	216	0.14	0.92	2
13	2.25	0.16	-1.29	209.63	216	8.06	0.01 [b]	2
14	-0.76	0.19	1.71	209.63	216	7.67	0.05	2
15	-1.38	0.23	-1.31	209.63	216	4.84	0.08	2
16	0.87	0.15	-2.17	209.63	216	6.00	0.05	2
17	0.44	0.15	-0.67	209.63	216	4.82	0.08	2
18	-1.05	0.21	1.35	209.63	216	8.53	0.01	2
19	-2.71	0.39	-0.79	209.63	216	1.18	0.55	2
20	2.59	0.17	-0.36	209.63	216	11.10	0.00 [b]	2
21	-1.09	0.21	-0.41	209.63	216	1.28	0.52	2
22	0.29	0.15	3.32 [a]	209.63	216	34.80	0.00 [b]	2
23	-2.98	0.45	-0.57	209.63	216	2.25	0.32	2
24	-0.33	0.17	0.22	209.63	216	1.16	0.55	2
25	0.96	0.15	-0.52	209.63	216	0.92	0.63	2
26	1.25	0.15	1.94	209.63	216	1.69	0.42	2
27	-0.16	0.16	1.38	209.63	216	6.26	0.05	2
28	-0.13	0.16	1.68	209.63	216	2.53	0.28	2
29	1.89	0.15	-1.64	209.63	216	2.19	0.33	2
30	0.77	0.15	1.66	209.63	216	5.02	0.08	2
31	-1.42	0.23	1.13	209.63	216	4.68	0.09	2
32	0.52	0.15	-2.04	209.63	216	3.46	0.17	2
33	-0.20	0.17	-0.31	209.63	216	1.29	0.52	2
34	1.48	0.15	-1.49	209.63	216	11.84	0.00 [b]	2
35	1.51	0.15	-2.37	209.63	216	11.33	0.00 [b]	2
36	0.63	0.15	3.86 [a]	209.63	216	12.69	0.00 [b]	2
37	0.89	0.15	0.24	209.63	216	2.87	0.23	2
38	3.10	0.19	1.17	209.63	216	10.22	0.01 [b]	2
39	1.50	0.15	-1.13	209.63	216	3.45	0.17	2
40	-2.52	0.36	-0.93	209.63	216	2.35	0.30	2

[a] denotes residual>3.0 – 3 items

[b] denotes p<0.05 – 7 items

Interestingly, the ANOVA results presented in Table 3 for some of the seven previously identified items provided evidence of the items not eliciting pre and post-treatment data that were statistically different or different in the same direction. This apparent concurrence between the parametric and stochastic test results requires comment. First, the parametric ANOVA contrasted

data from the two administrations whereas the Rasch analysis tested combined scores. Second, the parametric ANOVA comapred only mean test scores and the variance in these scores and did not examine conformity and separation between person ability and item difficulty parameters as did the Rasch analysis. The salient issues that arise from these observations are firstly the measurement capacity of the test and secondly, whether this could be improved prior to analysing the data to test for pre and post-treatment differences. While the ANOVA results suggest some of the items were contributing to errors in the aggregated data, comparing scale mean scores and variance between pre and post data does not adequately inform decisions about how the test could be improved. Alternatively, the Rasch analysis, particularly the individual item-fit statistics for the combined data does provide a strong basis for instrument refinement prior to testing for pre and post-treatment differences.

The content of items 4, 20, 22, 34, 35, 36 and 38 were carefully examined to ascertain whether their deletion would significantly diminish the content validity of the test. As these items concerned the ability to recognise the 'correct' English meaning of only a particular seven of the 40 Korean loanwords used in the test, the vocabulary range of the other 33 loanwords was considered sufficient. Consequently, data from the seven items were deleted from further analysis. The retained 33 items are presented in Appendix 1.

RUMM summary test-of-fit statistics were calculated for the 33-item data (see Table 5). The item-person interaction statistics revealed that for the combined pre and post-treatment data, the variance in item difficulty logits and student ability logits were similar (SD = 1.36 & 1.27 respectively). The student ability logits were greater than the item difficulty logits (Mean = 1.59 & 0.00 respectively) and this suggests some of the items were too 'easy' for these students. When the data fits the model, the fit residual has a mean near zero and a standard deviation near 1. The item and person fit residual means and standard deviations presented in Table 5 evidence good data to model fit. The item-trait interaction chi-square probability of 0.0016 suggests the test was measuring a dominant rather than uni-dimensional trait. The proportion of observed variance considered true should be close to 1. For these data the proportion of observed variance considered true (a Cronbach Alpha) was 0.82. The power of the test-of-fit statistic showed the overall fit between the data and the model was good with a separation index of 0.82.

Table 5
Display: SUMMARY TEST-OF-FIT STATISTICS (33 items N = 108)

ITEM-PERSON INTERACTION

	ITEMS		PERSONS	
	Location	Fit Residual	Location	Fit Residual
Mean	0.00	0.13	1.59	-0.15
SD	1.36	1.33	1.27	0.80

Complete data DF = 0.965

ITEM-TRAIT INTERACTION		RELIABILITY INDICES	
Total Item Chi Squ	104.97	Separation Index	0.82
Total Deg of Freedom	66.00	Cronbach Alpha	0.82
Total Chi Squ Prob	0.0016		

POWER OF TEST-OF-FIT
Power is GOOD
[Based on SepIndex of 0.82]

In general, the summary test-of-fit statistics show that the pre and post-treatment data from the 33 retained items could be plotted on one interval-level scale. This feature of the data enables accurate comparison of pre and post-treatment data.

Implicit in the decision to plot both pre and post-treatment data on one interval-level scale was the need to reduce errors that might occur if the data from the two administrations of the test were correlated. That is, pre-treatment test errors and post-treatment errors would be cumulative in a correlational analysis. Alternatively, when the combined data has been plotted on a single scale, the differences between pre and post-treatment scores can be estimated with a high level of certainty. To test for differences between pre and post scores for each of the 33 items, RUMM differential item functioning (DIF) was used.

When the raw data were entered into RUMM, a person factor design was specified in which the pre and post-treatment data were respectively identified by numeric codes. RUMM was instructed to generate Item Characteristic Curves including ANOVA statistics for the person factor (independent variable) of pre or post-treatment test for each item. Table 6 (below) presents the results of the ANOVA for Item 31 in the 33-item test.

The between-test F-ratio and probability values (0.09 & 0.77) suggest the independent variable of the test administration period has not significantly affected the dependent variable of the test score. This is because when RUMM classified the student ability (person location logits) into to three class intervals, the Mean of Observed Scores results showed different levels and opposite levels of the change in mean scores between the pre and post-treatment scores. For Class Interval 1 the mean score increased from 0.27 to 0.54, for Class Interval 2 it deceased from 0.48 to 0.42, and for Class Interval 3 it decreased from 0.80 to 0.77.

Table 6
Display: MEANS for ITEM 31

SOURCE	S.S	M.S	DF	F-RATIO	Prob
BETWEEN	5.05	1.01	5		
test	0.09	0.09	1	0.09	0.77
ClassInterval	0.89	0.45	2	0.42	0.65
test-by-CInt	4.06	2.03	2	1.90	0.15
WITHIN	221.59	1.07	207		
TOTAL	226.64	1.07	212		

Mean of Observed Scores for ITEM 31

Level	Total [N]	Level Mean	CInt1 Mean	CInt2 Mean	CInt3 Mean
Pretest	[108]	0.43	0.27	0.48	0.80
Posttest	[108]	0.66	0.54	0.42	0.77
Overall		0.54	0.32	0.46	0.78

Mean LOCATIONS for Class Interval [CInt]

Pretest	0.27	1.20	2.34
Posttest	0.32	1.19	2.90

The Item Characteristic Curve for Item 31 (labelled 37 according to the original instrument) presented in Figure 1 provides a graphical representation of the differences between the data from three class intervals of students. The expected posttest (post-treatment) scores were higher than the expected pretest scores for the students with lowest ability (class interval 1), but the direction of the difference was reversed for the two class intervals comprising the more capable students.

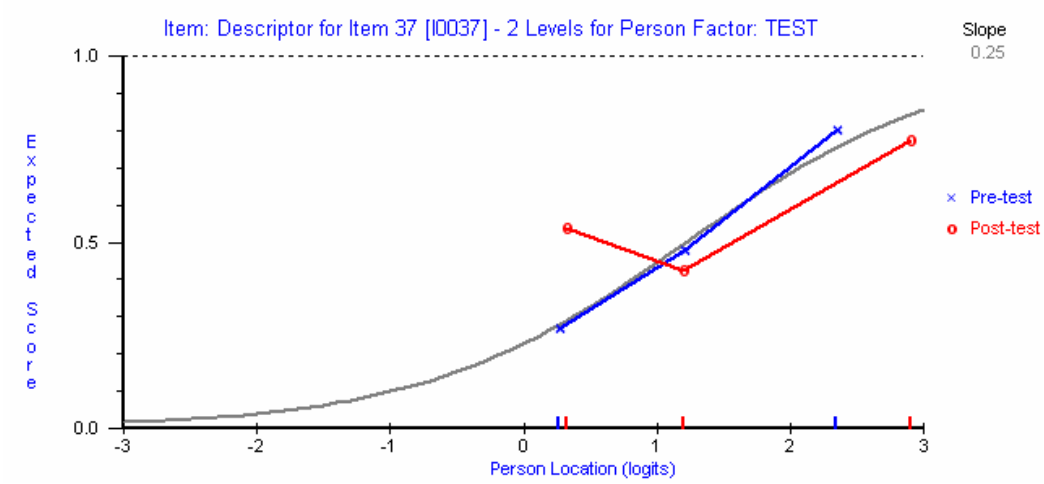

Figure1: RUMM plot of pre and posttest data – ANNOVA F-ratio p>0.05

Table 7 and Figure 2 present RUMM DIF output for Item 12 and this shows statistically significant changes in the dependent variable. The probability value for the between-test ANOVA is less than 0.05 (p = 0.01) and the mean observed scores have increased from the pretest to the posttest for the three class intervals of students. This is reflected in the Item Characteristic Curve that shows increasing levels of difference in the expected scores for the three classes of students.

Table 7
Display: MEANS for ITEM 12

SOURCE	S.S	M.S	DF	F-RATIO	Prob
BETWEEN	12.47	2.49	5		
test	7.33	7.33	1	8.64	0.01
ClassInterval	2.29	1.15	2	1.35	0.26
test-by-CInt	2.85	1.43	2	1.68	0.19
WITHIN	175.58	0.85	207		
TOTAL	188.05	0.89	212		

Mean of Observed Scores for ITEM 12

Level	Total [N]	Level Mean	CInt1 Mean	CInt2 Mean	CInt3 Mean
Pretest	[108]	0.09	0.06	0.11	0.13
Posttest	[108]	0.49	0.08	0.19	0.68
Overall		0.29	0.06	0.14	0.58

Mean LOCATIONS for Class Interval [CInt]

Pretest	0.27	1.20	2.34
Posttest	0.32	1.19	2.90

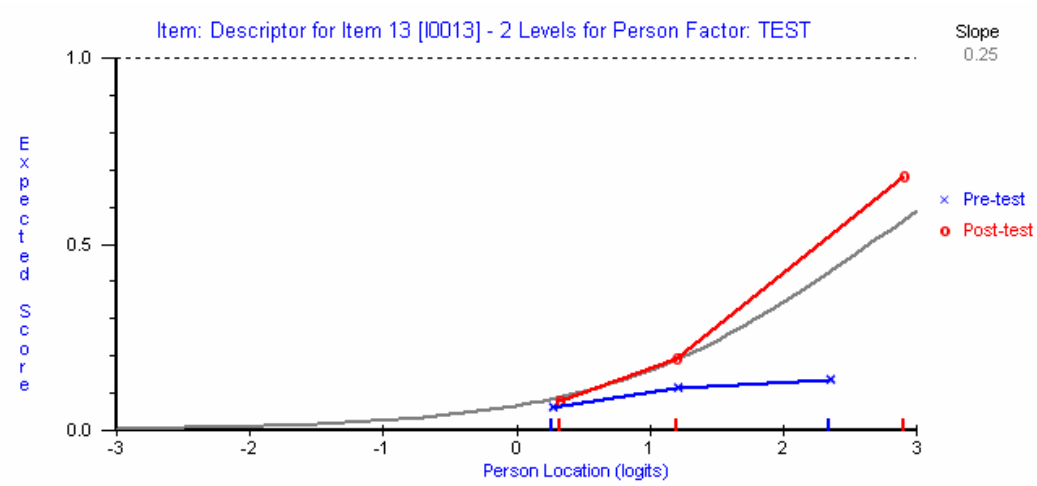

Figure 2: RUMM plot of pre and posttest data – ANNOVA F-ratio p<0.05

The RUMM DIF analysis results have been summarised in Table 8 for the 33-item data. This presents the between-test F-ratios and probability values for the 33-item pre and post-treatment data. Of the 33 items, only six showed a statistically significant ($p<0.05$) difference between the pre and post-treatment data.

Table 8
RUMM ANOVA (DIF) by pre and posttest transformed scores

Item*	Between-test F-ratio	Between test Prob
1	0.07	0.79
2	0.45	0.50
3	0.01	0.90
5	1.71	0.19
6	5.72	0.02 [a]
7	3.08	0.08
8	0.22	0.64
9	0.55	0.46
10	0.14	0.71
11	0.20	0.65
12	0.16	0.69
13	8.64	0.00 [a]
14	6.64	0.01 [a]
15	0.39	0.53
16	0.00	0.98
17	1.27	0.26
18	4.21	0.04 [a]
19	0.02	0.88
20	2.16	0.14
21	2.16	0.14
23	0.69	0.41
24	2.55	0.11
25	0.32	0.57
26	3.49	0.06
27	1.74	0.19
28	0.26	0.61

Table 8 continued		
29	0.20	0.04 [a]
30	0.01	0.93
31	0.89	0.35
32	0.72	0.40
33	0.47	0.50
37	0.09	0.77
39	6.20	0.01 [a]
40	1.32	0.25

* Items labelled according to the original 40-item test

[a] denotes $p<0.05$ – six items

Summary

The paired sample *t*-test of pre and posttest raw scores for the 40-item test (see Table 1), suggests there were statistically significant ($p<0.01$) differences between the pre and posttest scores. If this test and the use raw scores are considered valid, either the treatment independent variable or possibly an extraneous variable(s) has affected student performance on the dependent variable. Similarly, the ANOVA results of the raw scores for the 40-item test (see Table 2) could be taken as evidence of the strong effect of the independent variable or possibly an extraneous variable(s) on the dependent variable. This finding is strengthened by the high effect size statistic ($eta^2 = 0.31$).

The item by item ANOVA results presented in Table 3 revealed a more complex relationship between the pre and posttest raw scores. As was noted previously, these results raise questions about the internal reliability and content validity of the 40-item test. While there was evidence of a need to refine the test, the use of parametric statistics did not provide sufficiently detailed information to enable this to be undertaken with a high degree of certainty.

Application of the stochastic Rasch model in the RUMM analyses (see tables 4 to 7 and Figures 1 & 2), could be viewed as a more appropriate method for the following reasons. First, both student ability and item difficulty parameters were taken into account. Second, student ability and item difficulty logits were plotted on one interval-level scale. Third, pre and post-treatment data were plotted on the same interval-level scale which enabled an accurate comparison of these data by reducing errors of measurement. Fourth, the effects of covariates or extraneous variables were partially controlled by plotting pre and post-treatment data on one interval-level scale and not considering data with poor fit to the measurement model. It is likely that the poor data to model fit from the seven deleted items was due to the inconsistent effects of extraneous variables on pre and post-treatment student ability. However, the omission of a control group and random assignment of subjects in the research design did limit the extent to which extraneous variables were controlled.

Notwithstanding the data having been obtained from a one-group pretest - posttest investigation, the comparison of deterministic and stochastic data analysis techniques has shown the benefits of applying stochastic procedures. The central issue in this assertion is the requirement for interval data in parametric analyses and being sure the data has this psychometric property.

Conclusion

The problem investigated in this report concerned the application of detrministic and probabilistic analyses in experimental research and the extent to which the respective results provide confidence for confirming or rejecting the inherent research hypothesis. An examination of the paired sample *t*-test of raw scores, ANOVA of raw scores, and ANOVA of Rasch transformed scores identified differences in the techniques that could influence the accuracy of the respective tests. When the

same data were analysed by the three methods, the results provided varying levels of evidence for hypothesis confirmation.

For aggregated raw scores, the *t*-test and ANOVA results showed a statistically significant increase in scores from the pretest to posttest. Conducting ANOVA of the raw scores for individual items suggested that the aggregation of the raw scores had probably masked differences in item reliability and item content validity. Investigation of this potential scale deficiency was not possible using parametric methods so a Rasch analysis was conducted. This showed that the measurement capacity of the scale could be improved by deleting items that were contributing to errors of measurement in the data.

A second Rasch analysis of the refined scale including ANOVA of transformed scores revealed that six of 33 items elicited data that were statistically different between the pretest and posttest. This finding provides only a paucity of evidence for confirming the research hypothesis.

In conclusion, the study has illustrated why parametric analysis using non-interval data should be undertaken with caution in experimental research. Further, it has demonstrated the benefits of applying a probabilistic data analysis technique in this type of research.

References

Andrich, D., Sheridan, B., Lyne, A., & Luo, G. (2000). *RUMM: a windows-based item analysis program employing Rasch unidimensional measurement models*. Perth: Murdoch University.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Daulton, F. E. (1998). Japanese loanword cognates and the acquisition of English vocabulary. *The Language Teacher Online, 22(1).* Retrieved: July 14, 2003, from: http://langue.hyper.chuba.ac.jp/jalt/pub/tlt/98/jan/daulton.html.

Daulton, F. E. (1999a). English loanwords in Japanese – The Built in Lexicon. *Internet TESOL Journal.* 5(1). Retrieved August 23, 1999 from: http://www.aitech.ac.jp/~iteslj/Articles/Daulton-Loanwords.html.

Daulton, F. E. (2003). List of high-frequency baseword vocabulary for Japanese students #2. *Internet TESOL Journal, IX(3).* Retrieved July 14, 2003, from: http://iteslj.org/lists/Daulton-BasewordVocab2.html.

Elgort, I., Marshall, S., & Mitchell, G. (2003). NESB Student perceptions and attitudes to a new online learning environment. *The Higher Education Research and Development Society of Australasia Conference, July 6-9 2003. Christchurch: New Zealand*. Retrieved July 26, 2003, from: http://surveys.canterbury.ac.nz/herdsa03/pdfsref/Y1049.pdf

Fraenkel, J.R., & Wallen, N.E. (2004). *How to design and evaluate research in education*. New York: McGraw Hill.

Kang, S. M. (2000). Interactive multimedia: A pilot study in using homepage and email in English class. *Multimedia Assisted Language Learning, 3(1), 47-66*.

Keem, S. U. (2000). A field study: Multimedia assisted English instruction to cultivate communicative competence. *Multimedia Assisted Language Learning, 3(1), 139-166*.

Kent, D. B. (1996). *Kent Konglish dictionary,* [Software Download]. Kent Interactive Labs. Retrieved December 20, 1996, from: http://www.geocities.com/Tokyo/Towers/5067/kkd.html.

Kent, D. B. (2001). Teaching Konglish: Selected resources for students and teachers. *Korean Teachers of English to Speakers of Other Languages Conference*. October 13-14 2001. The Learning Environment: The Classroom and Beyond. Seoul, Korea.

K.E.R.I.S. (2001). *Adapting education to the information age: A white paper*. Ministry of Education and Human Resources Development, Korean Education and Research Information Service. Retrieved January 27, 2003, from http://www.keris.or.kr/english/2001-WhitePap.pdf.

Kiess, H.O. (1996). *Statistical concepts for the behavioural sciences*. Sydney: Allyn and Bacon.

Lee, C. I., & Yang, E. M. (2002). Integrating CALL into classroom practices: Its theory and application. *The Korean Association of Multimedia Assisted Language Learning International Conference Proceedings, 169-183.* A New Paradigm for Innovative Multimedia Language Education in the 21st Century. October 3-5, 2002. Seoul, S. Korea.

Min, S. J., Kim, H. K., & Jung, K. T. (2000). A paradigm shift in English education in Korea: Integration of the textbook to a web-based curriculum. *Multimedia Assisted Language Learning, 3(1), 119-138.*

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition. Chicago: The University of Chicago Press, 1980.)

Romanoski, J., & Douglas, G. (2002). Test scores, measurement, and the sse of analysis of variance: An historical overview. *Journal of Applied Measurement, 3(3), 232-242.*

Shaffer, D. (2001). A New approach for a new language. *The Internet TEFL Journal, 35.* Retrieved July 26, 2003, from: http://www.mantoman.co.kr/issues/m035/m3510.htm.

Shaffer, David. (1999). *Picture that! – drawing techniques for teaching false cognates*. Second Pan-Asian International Conference. Teaching English: Asian Contexts and Cultures. October 1-3. Seoul, S. Korea.

Sheperd, J. W. (1996). Loanwords a pitfall for all students. *Internet TESOL Journal 2(2), 1-8*. Retrieved February 17, 1996 from: http://www.aitech.ac.jp/~iteslj/Articles/Shepherd-Loanwords.html.

Simon-Maeda, A. (1997) Language awareness: Use/misuse of loan-words in the English language in Japan. *Internet TESOL Journal, I(2)*. Retrieved May 19, 1997, from: http://www.aitech.ac.jp/~iteslj/Articles/Maeda-Loanwords.html.

SPSS. (2003). *SPSS 12.01 for Windows*. Chicago: SPSS Inc.

Taylor, I. S., &Taylor, M. M. (1995). *Writing and literacy in Chinese, Korean and Japanese*. Philadelphia: John Benjamins Publishing Company.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis: Rasch measurement*. Chicago. IL, MESA.

Appendix 1: The test

QUIZ

▪ ▪ ▪ ▪

▪ ▪ ▪ ▪ (UE106-26)

▪ ▪ ▪ ▪ (▪ ▪ ▪ ▪ ▪ ▪)

……
…
……
…
……
…

TEST INSTRUCTIONS

아래 나온 각각의 문장 속에 밑줄로 쳐진 단어가 있다. 각각의 단어에 대해 제시된 사지 선택 중에서, 답으로 가장 알맞은 영어 정의를 고른다.

In each of the sentences below there is an underlined term(s), this term is a 'pseudo loanword'. From the four choices provided for each term, select the definition that equals the English meaning for the underlined term.

1.	Your neighbor's audio is really loud, someone should complain to the police.		
Definition Choices			
a (1a)	shirt	**c** (1c)	noise
b (1b)	dog	**d** (1d)	stereo

2.	Kangnam is always a good area in Seoul to go hunting.		
Definition Choices			
a (2a)	sleeping	**c** (2c)	camping
b (2b)	picking up people	**d** (2d)	eating

3.	We need to get new batteries for the remocon it doesn't seem to be working anymore.		
Definition Choices			
a (4a)	brand of batteries	**c** (4c)	remote control
b (4b)	ready mixed concrete	**d** (4d)	a reverse cycle air conditioner

5.	What night do you like the best?		
Definition Choices			
a (7a)	night club	**c** (7c)	bar
b (7b)	saturday	**d** (7d)	the opposite to day

6.	When she fell over, she landed on her hip.		
Definition Choices			
a (8a)	front	**c** (8c)	top
b (8b)	bottom	**d** (8d)	back

7.	It is really hot in here, is there something wrong with the steam?		
Definition Choices			
a (10a)	an angry person	**c** (10c)	water
b (10b)	engine	**d** (10d)	radiator heater

8.	The sharp is not in your pencil case.		
Definition Choices			
a (11a)	pen that never runs out of ink	**c** (11c)	pointed paper
b (11b)	mechanical pencil	**d** (11d)	knife that can cut well

9.	Come in, and take a seat on the sofa over there.		
Definition Choices			
a (12a)	train station seat	**c** (12c)	armchair
b (12b)	love seat	**d** (12d)	place that is not near here

10.	I couldn't see anything because I needed a flash.		
Definition Choices			
a (13a)	flashlight	**c** (13c)	the sound of thunder after lightening
b (13b)	fresh batteries	**4** (13d)	to move like a turtle

11.	I had to take good hold of the handle before I could get the car to turn properly.		
Definition Choices			
a (14a)	part of a machine used to turn it on and off	**c** (14c)	steering wheel
b (14b)	a persons ability to control a machine	**d** (14d)	the end of the arm

12.	Why do so many people blow their klaxon when driving in Korea?		
Definition Choices			
a (15a)	car horn	**c** (15c)	boat horn
b (15b)	siren	**d** (15d)	alarm

13.	He is a talent that is really very well known.		
Definition Choices			
a (16a)	a singer	**c** (16c)	an actor or actress
b (16b)	a media celebrity	**d** (16d)	an expert

14.	Your hands look really nice, you usually don't use manicure.		
Definition Choices			
a (17a)	nail polish	**c** (17c)	hand treatment
b (17b)	hand cream	**d** (17d)	cut and trimmed toenails

15.	I don't like the potato in this store, let's go to another place.		
Definition Choices			
a (18a)	french fries	**c** (18c)	toes
b (18b)	pots	**d** (18d)	hash browns

16.	Oh no, another mistake. I can't see why I just did that. Where's the white?		
Definition Choices			
a (19a)	ball point pen	**c** (19c)	the person a man marries
b (19b)	eraser	**d** (19d)	correction fluid

17.	She always goes on a meeting.		
Definition Choices			
a (21a)	blind date	**c** (21c)	promise
b (21b)	lecture	**d** (21d)	timetable

18.	Don't write with a ball pen please use a pencil.		
Definition Choices			
a (22a)	sharp	**c** (22c)	fountain Pen
b (22b)	mechanical pencil	**d** (22d)	ballpoint Pen

19.	Every weekend my friends and I go out to a hof or two.		
Korean Term Choices			
a (24a)	amusement park	**c** (24c)	fast food chain
b (24b)	bar	**d** (24d)	cinema

21.	Plug this into the consent over there.		
Definition Choices			
a (26a)	hole	**c** (26c)	water outlet
b (26b)	electrical outlet	**d** (26d)	wall

23.	He's really handsome I can't believe he's still solo.		
Definition Choices			
a (28a)	a lemon flavored drink	**c** (28c)	single
b (28b)	short	**d** (28d)	married

24.	Everybody was fighting for their team at the soccer game last night.		
Definition Choices			
a (29a)	boxing	**c** (29c)	cheering
b (29b)	rioting	**d** (29d)	crying

25.	I'll need to buy a new spring note after winter because this one will be full.		
Definition Choices			
a (30a)	seasonal message	**c** (30c)	spring notebook
b (30b)	spiral bound notebook	**d** (30d)	spiral bound note

26.	I don't know why the coach hasn't called for a member change yet.		
Definition Choices			
a (31a)	time out	**c** (31c)	substitution
b (31b)	water boy	**d** (31d)	change of teams

27.	Be careful in this area otherwise you'll end up with a punk.		
Definition Choices			
a (32a)	flat tire	**c** (32c)	person with long hair
b (32b)	pumpkin	**d** (32d)	Damage

28.	I spend a lot of time training with the sand bag.		
Definition Choices			
a (33a)	plastic bag	**c** (33c)	bag full of sand
b (33b)	beach bag	**d** (33d)	punching bag

29.	That company has an excellent reputation for after service.		
Definition Choices			
a (35a)	after sales service	**c** (35c)	buying things after getting something free
b (35b)	expensive repairs	**d** (35d)	after selling service

30.	I like the sand type biscuits.		
Definition Choices			
a (36a)	gritty textured	**c** (36c)	fine grained
b (36b)	smooth	**d** (36d)	sandwich

31.	I'm just going to the super for a moment.		
Definition Choices			
a (37a)	street market	**c** (37c)	excellent
b (37b)	gym	**d** (37d)	supermarket

32.	Do you want to play a game of pocket ball?		
Definition Choices			
a (38a)	billiards	**c** (38c)	fashionable clothing
b (38b)	ball you keep in your pocket	**d** (38d)	game like basketball

33.	I'll need to get a driver before I can help you repair it.		
Definition Choices			
a (39a)	spanner	**c** (39c)	tool
b (39b)	screwdriver	**d** (39d)	shoe

b (44b)	underpants	**d** (44d)	outerwear

37.	Everybody one shot!		
Definition Choices			
a (45a)	bottom's up	**c** (45c)	taste
b (45b)	cheer	**d** (45d)	sip

39.	I drank too much, I think I'm going to o-bite.		
Definition Choices			
a (49a)	eat more	**c** (49c)	over bite
b (49b)	over eat	**d** (49d)	vomit

40.	Is your a-pa-teu nearby?		
Definition Choices			
a (50a)	pull apart	**c** (50c)	put together
b (50b)	house	**d** (50d)	apartment